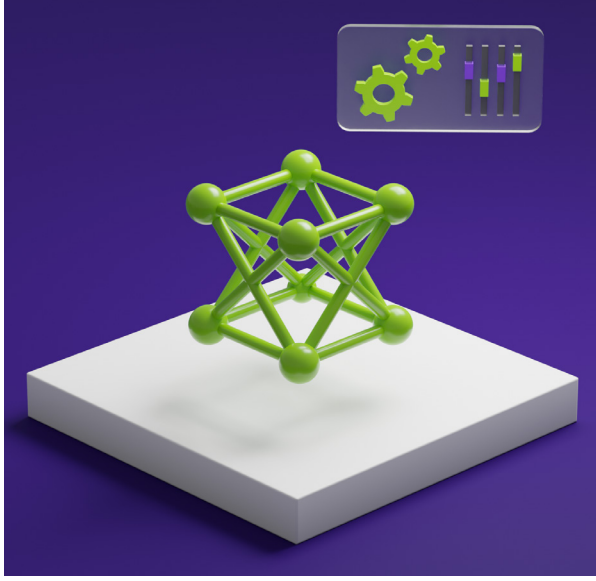
Datasheet

Efficient Large Language Model (LLM) Customization



Workshop Overview

Enterprises need to execute language-related tasks daily, such as text classification, content generation, sentiment analysis, and customer chat support, and they seek to do so in the most cost-effective way. Large language models can automate these tasks, and efficient LLM customization techniques can increase a model's capabilities and reduce the size of models required for use in enterprise applications.

In this course, you'll go beyond prompt engineering LLMs and learn a variety of techniques to efficiently customize pretrained LLMs for your specific use cases—without engaging in the computationally intensive and expensive process of pretraining your own model or fine-tuning a model's internal weights. Using NVIDIA NeMo™ service, you'll learn various parameter-efficient fine-tuning methods to customize LLM behavior for your organization.

Learning Objectives

By participating in this workshop, you'll learn how to:

- Apply parameter-efficient fine-tuning techniques with limited data to accomplish tasks specific to your use cases
- Use LLMs to create synthetic data in the service of fine-tuning smaller LLMs to perform a desired task
- Leverage NVIDIA NeMo service to customize models like GPT and LLaMA-2 with ease

Overview	
Duration	8 hours
Price	**Contact us for pricing**.
Prerequisites	> **Professional experience with the Python programming language** > **Familiarity with fundamental deep learning topics like model architecture, training, and inference** > **Familiarity with a modern Python-based deep learning framework (PyTorch preferred)** > **Familiarity working with out-of-the-box pretrained LLMs** > Experience with advanced prompt engineering techniques
Tools, libraries, and frameworks	Python, NVIDIA NeMo Service, GPT, LLaMA-2
Assessment type	Skills-based coding assessments evaluate the learner's ability to efficiently customize and compose pretrained LLMs.
Certificate	Upon successful completion of the assessment, participants will receive an NVIDIA DLI certificate to recognize their subject matter competency and support professional career growth.
Hardware requirements	Desktop or laptop computer capable of running the latest version of Chrome or Firefox. Each participant will be provided with dedicated access to a fully configured, GPU-accelerated workstation in the cloud.
Language	English

Workshop Outline	
Introduction (15 minutes)	Meet the instructor. > Create an account at **courses.nvidia.com/join**
Parameter-Efficient Fine-Tuning (PEFT) Essentials (115 minutes)	Investigate PEFT techniques like low-rank adaptation (LoRA) and p-tuning to tailor LLMs for specific tasks using limited data: > Grasp the principles of PEFT techniques like LoRA and p-tuning and how they efficiently fine-tune LLMs without direct modification to the LLM. > Learn how to acquire and prepare data for use in parameter-efficient fine-tuning. > Perform LoRA and p-tuning on a variety of GPT LLMs while quantitatively analyzing fine-tuned model performance.
Break (45 minutes)	
PEFT for Reduced Model Sizes (115 minutes)	Learn a set of techniques for using larger prompt-engineered LLMs to generate synthetic data, which can be used to create smaller parameter-efficient fine-tuned models capable of the same task: > Create LLM functionality specifically suited for common synthetic data generation tasks. > Perform LoRA using synthetic data from a larger model's responses to fine-tune a smaller model capable of the same task. > Create a virtuous cycle of using LLMs to create synthetic data to fine-tune smaller LLMs capable of creating more synthetic data.
Break (30 minutes)	
Engineering With Customized Models (115 minutes)	Practice engineering application code that combines multiple fine-tuned models: > Create multiple fine-tuned small LLMs for sentiment analysis, extractive question answering, text generation, and persona creation. > Learn techniques for creating LLMs capable of speaking in your own style of writing. > Compose multiple small fine-tuned LLMs into meaningful application code.
Assessment and Q&A (45 minutes)	Use the techniques from this workshop to create an application that uses several small fine-tuned LLMs to perform detailed analysis of customer emails and generate specific automatic responses.

Why Choose NVIDIA Deep Learning Institute for Hands-On Training?

> Access workshops from anywhere with just your desktop/laptop and an internet connection. Each participant will have access to a fully configured, GPU-accelerated server in the cloud.

> Obtain hands-on experience with the most widely used, industry-standard software, tools, and frameworks.

> Learn to build deep learning and accelerated computing applications for industries, such as healthcare, robotics, manufacturing, accelerated computing, and more.

> Gain real-world experience through content designed in collaboration with industry leaders, such as the Children's Hospital of Los Angeles, Mayo Clinic, and PwC.

> Earn an NVIDIA DLI certificate to demonstrate your subject matter competency and support your career growth.

Ready to Get Started?

For the latest DLI workshops and trainings, visit
www.nvidia.com/dli

For questions, contact us at **nvdli@nvidia.com**

NVIDIA